Journal of Biomedical Informatics 45 (2012) 231–239

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Summary of Product Characteristics content extraction for a safe drugs usage

S. Rubrichi *, S. Quaglini

Laboratory for Biomedical Informatics, "Mario Stefanelli", Department of Computers and Systems Science, University of Pavia, Pavia, Italy

ARTICLE INFO

Article history:
Received 27 March 2011
Accepted 30 October 2011
Available online 10 November 2011

Keywords:
Information extraction
Conditional random fields
Support vector machines
Adverse drug events
Medication errors
Summary of Product Characteristics

ABSTRACT

The use of medications has a central role in health care provision, yet on occasion, it may injure the person taking them as result of adverse drug events. A correct drug choice must be modulated to acknowledge both patients' status and drug-specific information. However, this information is locked in free-text and, as such, cannot be actively accessed and elaborated by computerized applications. The goal of this work lies in extracting content (active ingredient, interaction effects, etc.) from the Summary of Product Characteristics, focusing mainly on drug-related interactions, following a machine learning based approach. We compare two state of the art classifiers: conditional random fields with support vector machines. To this end, we introduce a corpus of 100 interaction sections, hand annotated with 13 labels that have been derived from a previously developed conceptual model. The results of our empirical analysis demonstrate that the two models perform well. They exhibit similar overall performance, with an overall accuracy of about 91%.

1. Introduction

Adverse drug events (ADEs) have been defined as injuries resulting from medical intervention related to a drug [1] that endanger patient's safety and account for increased heath care costs. Examples include injuries (e.g., rash, confusion, or loss of function) caused by incorrect dosage as well as allergic reactions occurring in a patient not known to be allergic to a given medication and so forth. Many of these injuries suffered by patients are inevitable but at least a quarter may be secondary to medication errors [2], as errors in medication-management process are generally called [3]. These damages are not unavoidable and can be prevented. According to one estimate, medication errors occur most frequently at the prescribing and administration stages [2]. The rate of medication errors and preventable ADEs represents a serious cause for concern. The Institute of Medicine (IOM) committee, in its report Preventing Medication Error, estimates that more than 1.5 million ADEs are preventable each year in the US alone [2]. This report outlines the main priorities for research on safe medication use that will address this problem. It proposes electronic prescribing, by computerized provider order entry (CPOE) systems, as one of the most highly effective error prevention strategies.

CPOE systems are computer applications that allow direct, electronic entry of orders for medications, laboratory, radiology, referral, and procedures [4]. Several systematic reviews have shown the benefits of CPOE systems [5–10] resulting from their ability to detect unsafe and potentially fatal medication orders. Computerization in fact enables the delivery of clinical decision support [11], alerts to guide ordering, allowing for checks for allergies, drug–drug interactions, clinical conditions. One of the factors, which mainly influences the overall performance of such a system, is the quality and validity of the knowledge base underlying the system. Safe medication use requires that providers and consumers synthesize several types of information, including knowledge of the medication itself, as well as understanding of how it may interact with coexisting illnesses and medications, and how its use might be monitored. This information is constantly changing, and while most of the necessary updated knowledge is available somewhere, it is not always readily accessible, creating a situation in which it is almost impossible for health care providers to have current knowledge of every medication they prescribe.

In this work, we consider the problem of automatic extraction of drug information conveyed in the Summary of Product Characteristics (SPC), focusing on a specific section concerned with drug-related interactions. Our contributions are as follows:

1. We formulate the problem in a machine learning framework, in which we seek to assign the correct semantic label, such as *InteractionEffect* or *ActiveDrugIngredient*, to each word, or segment of sentence, of the text. We employ two state of the art classifiers: linear-chain conditional random fields (CRFs) and structural support vector machines (SVMs). These classifiers discriminate between semantically interesting and uninteresting content through the automatic adaptation of hundreds of engineered text characteristics, taking into account the properties of a document, on both a local (word) and global (sentence) level.

* Corresponding author. Fax: +39 0382 525638.
E-mail address: stefania.rubrichi@unipv.it (S. Rubrichi).

1532-0464/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.10.012

2. We introduce a corpus of 100 interactions sections in Italian language that have been annotated with 13 distinct semantic labels, with respect to a previously implemented ontology.
3. We apply the CRFs and the SVMs to our data set and evaluate their overall and individual label results. Both the classifiers achieve a micro-averaged F_1-measure (see Section 4.2) greater than 90%, which is promising for real-world applications.

In a next step, the extracted information will be used to populate the ontology, in order to carry out automated reasoning on data. This reasoning process could help, for example, to determine whether it is possible for a particular drug to have any interaction with a particular active ingredient or diagnostic test, to find the effect of a particular interaction and so on. Moreover, ontology population can compensate for the possible lack of completeness and/or congruence among different SPCs. Like this, such knowledge model can be made available to specific prescription applications, such as CPOE, for integrating the underlying base of knowledge, thus improving the prescription process.

2. Background

SPCs represent a vast source of information for health professionals on how to use medicines safely and effectively. It forms an intrinsic and integral part of marketing authorization. In order to obtain an authorization to place a medicinal product on the market, a SPC shall be included in the application made to the competent authority. A SPC lays out the agreed position (results of physico-chemical, biological or microbiological tests, toxicological and pharmacological tests, clinical trials, etc.) on the medicinal product as collected during the course of the assessment process. Its content is regulated by Article 11 of Directive 2001/83/EC. Accordingly, SPCs of specialty medicines for human use are organized into 12 sections: name, therapeutic categories, active ingredient, excipients, indications, contraindication/side effects, undesired effects, posology, storage precautions, warnings, interactions, and use in case of pregnancy and nursing.

All this information is locked in free-text, the most convenient and natural way to convey medical knowledge for human communication [12]. This narrative form, however, cannot be actively used by health information systems. Reliable access to this comprehensive information, by Natural Language Processing (NLP) systems, can provide a wide range of coded data [12,13]. Such data will then be available for new or enriched clinical applications, thus facilitating and improving the prescription process. It is therefore an important aspect for improving quality of care and preventing medication errors.

Among NLP techniques, information extraction (IE) methods have been largely employed in the biomedical domain to extract facts from free-text [14–18] and to make them available for subsequent tasks such as case finding, summarization, decision support, or statistical analysis.

In this work, we therefore propose to extract drug-related interaction information reported in SPC, following a named entity recognition (NER) approach. NER is an important step in an integral IE task and aims at identifying words or phrases in natural language text belonging to certain classes of interest (i.e., Named Entities), such as diseases or drugs, and labeling them with their appropriate type [19]. In NER, an attempt will be made to associate each token with a label that indicates its appropriate domain-specific category. Approaches to NER span a broad range, from rule-based systems to machine learning. Rule-based systems make decisions on sets of hand-written disambiguation rules that play an important role in discovering a named entity, which specify, for example, that an ambiguous word belongs to a particular named entity rather than to another one if it follows another specific named entity.

On the other hand, a typical application of machine learning works to classify a novel instance x as belonging to a particular class y. In the field of label sequence problems such techniques aim at identifying the most likely sequence of labels for the words in any given sentences. These methods generally resolve tagging ambiguities by using training corpus to compute the probability of a given word having a given tag in a given context. Then, they automatically tune their own parameters to maximize their performance on the training corpus. The machine then generalizes from these samples.

Rule-based NER systems can be very effective but require some manual effort. Machine learning approaches can successfully extract named entities but require large annotated training data. Advantages of machine learning approaches are that they do not require human intuition, they are general and clearly separate the algorithm from the data, so that it is easier to apply them to any domain by simply retraining without reprogramming.

2.1. Related work

Several studies have addressed the issue of IE in medication domain. Some approaches concentrated their analysis on the extraction of drug names. Levin et al. [20] implemented a system based on lexicon (RxNorm) and regular expressions (Hints List) to extract and normalize drug names from an anesthesia electronic health record, into a standardized terminology. RxNorm and Hints List concepts were used in the mapping module as references for drug names, and medical abbreviations and jargon, respectively. In another study, Sirohi and Peissig [21] performed a dictionary-based NLP study to determine the effects of using varying lexicon to extract drug names from electronic medical records. These authors have shown how the accuracy of results can be enhanced by refining the drug lexicon.

Other studies focus on extracting more specific drug features, such as drug names and dosage. In one example, Evans et al. [22] reported a method of extracting drug and dosage data from a collection of discharge summaries. They first draw a conceptual model of drug-dosage information and then identified this information using a semantically driven extraction module. This module combines readily available NLP facilities from the Clarit system with newly created resources, including a set of pattern rules and a lexicon. A study by Shah and Martinez [23] derived numerical information about daily dosage from unstructured dosage instructions from a patient records database, using a dictionary to standardize words and phrases. Then, they converted the extracted information into structured fields.

Lately, more studies have been geared toward the extraction of a more complete set of drug characteristics. In particular, Gold et al. [24] built Merki, a parser that can extract drug names and other relevant information from discharge summaries using a lexicon and a set of parsing rules. Evaluation showed that the system identified drug names, but other information such as dose and frequency had lower precisions. Similarly, Xu et al. [25] implemented a NLP system, MedEx, which extracts medication information from clinical notes. Relying on a more detailed medication representation model, they integrated a semantic tagger and Chart parser to capture drug names and signature information from clinical narratives and then to map it onto structured representation.

The Third i2b2 NLP Challenge [26] focused on the extraction of medication information, in particular medication names, dosages, routes of administration, frequencies, durations, and reasons for administration, from discharge summaries. Different approaches have been proposed to address this task: rule-based, machine learning, and hybrid systems. Among others, Doan et al. [27] integrated

different existing NLP components that included a sentence boundary detection program, a section identification program (SecTag), the MedEx system for tagging medication fields, and finally a context-free grammar for converting the text into a structured format. Tikk and Solt [28] proposed an approach in three steps: NER, context filtering, and relation extraction. While the last two components are rule-based, for NER, they investigated first a rule-based and second a CRFs-based method, which, with enough large training data, showed considerably better performance. Meystre et al. [29] employed a hybrid system, between machine learning and rule-based, called Textractor.

On the whole, current works focus on clinical text narrative. However, Pereira et al. [30] considered another source of information on medications, namely SPCs, and thus addressed the problem of automatic indexing. The authors developed a method to automatically generate a dictionary for use with a French Multi-Terminology Indexing tool.

Only a few published papers address the issue of extracting drug interactions information from narrative text. Segura-Bedmar et al. [31] employed a kernel-based approach that uses SVMs aiming at detecting drug–drug interactions. A more detailed analysis of such work and a comparison with our own are presented in Section 5.

3. Methods

As we have already stated, we developed a framework for simultaneously recognizing occurrences of multiple entity classes using linear-chain CRFs and structured SVMs. Both of these two supervised machine learning approaches predict words' labels by using a large number of interdependent descriptive characteristics (features) of the input by assigning real-valued weight to these features.

Presented here is an outline of our framework. We propose an approach in five steps. We began by defining a semantic representation model of drug information conveyed in the SPCs, in order to find out the concept to be extracted. In a second step, a preprocessing pass was required for preparing the dataset for the use by the extraction module. Then, we annotated the text by hand, with respect to the previously developed conceptual model. Subsequently, we defined a set of binary features that express some descriptive characteristics of the data, for instance "current token is capitalized". Finally, we processed the data through the two discriminative models (CRFs and SVMs): both the algorithms iterate the tokens in the sentence, and label proper tokens with semantic labels, by learning the correspondence between labels and features.

3.1. Conceptual model

Typically, the first step in most NER tasks is to identify the named entities (labels) that are relevant to the concepts, relations, and events described in the text. A system for NER is thus based upon specific knowledge about the domain. Therefore, as part of the understanding of the text factual information process, we had previously developed a formal model of drug information conveyed in the SPCs. We conducted a manual analysis of SPCs text so as to identify the underlying semantic concept classes (i.e., concepts representing drug features) and semantic relationships among those concepts. This analysis has resulted in a domain ontology of medication [32], the formal means of representing domain-specific knowledge. In particular, in this study, we focused on drug interactions, then we looked, more specifically, at the 12 concepts that properly model drug interaction findings. A partial view of the implemented ontology, concerned with drug-related interactions, is presented in Fig. 1.

Generally, a drug interaction represents the situation in which a substance affects the activity of an active ingredient, resulting in various effects such as alterations in absorption, metabolism, excretion, and pharmacodynamics (i.e., the drug effects are decreased or increased, or the drug produces a new effect that neither produces on its own). Typically, interaction between active ingredients comes to mind. However, interactions may also exist between drugs and foods, as well as drugs and herbs. Moreover, some interactions can be abstracted at a drug class level. Eventually, an interaction can occur under particular contexts, and in the presence of particular treatment conditions (such as dosage, intake route). All this information is made explicit in the ontology by the relations "with", "produces" and "under condition". The first links the class "Interaction" to the classes "Active Drug Ingredient", "Drug Class" and "Other Substance" to define the substance interfering. The second links the class "Interaction" to the classes "Interaction Effect", while the latter links "Interaction" with "Personal Condition", "Posology" and "Intake Route", in order to define the conditions potentially leading to the occurrence of the interaction effect.

3.2. Preprocessing step

As long as the label prediction is on a word-by-word basis, and decisions are made for one sentence at a time, the first stage of our extraction algorithm consists in splitting the text of SPC interaction section into sentences and then to break those input sentences into tokens. We used full stops and white spaces to determine sentence and token boundaries, respectively. Moreover, in order to account for exceptions, we considered a normalization steps that mainly includes removing all punctuation but colon and brackets, adding white spaces between colon and brackets, and the previous word, removing hyphens if they exist between alphanumeric strings, replacing dots that occur between numbers as decimal mark, ("4.5") with commas ("4,5").

3.3. Annotation process

The annotation process was performed by a biomedical engineer with domain knowledge. Semantic annotation is used to establish links between the entities in the text and their semantic descriptions or concept classes in the above described ontology. We used the following 13 semantic labels: *ActiveDrugIngredient, AgeClass, ClinicalCondition, DiagnosticTest, DrugClass, IntakeRoute, OtherSubstance, InteractionEffect, Posology, PharmaceuticalForm, PhysiologicCondition, RecoveringAction, None*. The latter has been given to indicate elements that are not relevant for this research.

Leveraging the established ontology, we mapped its elements to the SPCs' text content. We carefully inspected all the corpus lines distinguishing the different senses with respect to the ontology; we then annotated each word in the extracted SPC interaction sections with the corresponding class in the ontology, by assigning a HTML tag. Active ingredients tagging was performed automatically, by a look-up of terms in Farmadati active ingredients archive. In particular, we implemented a Java application which, for each active ingredient in the archive, tests if the text matches it and then adds the tag. A review of the data has been used to validate and, when necessary, correct the annotations.

As an example, consider the following sentence (translated from Italian):

⟨**Salicylates**⟩$_{DrugClass}$ ⟨**may enhance the effect**⟩$_{InteractionEffect}$ **of** ⟨**oral**⟩$_{IntakeRoute}$ ⟨**hypoglycemic agents**⟩$_{DrugClass}$, ⟨**eptifibatide**⟩$_{ActiveDrugIngredient}$ **and** ⟨**sodium valproate**⟩$_{ActiveDrugIngredient}$.

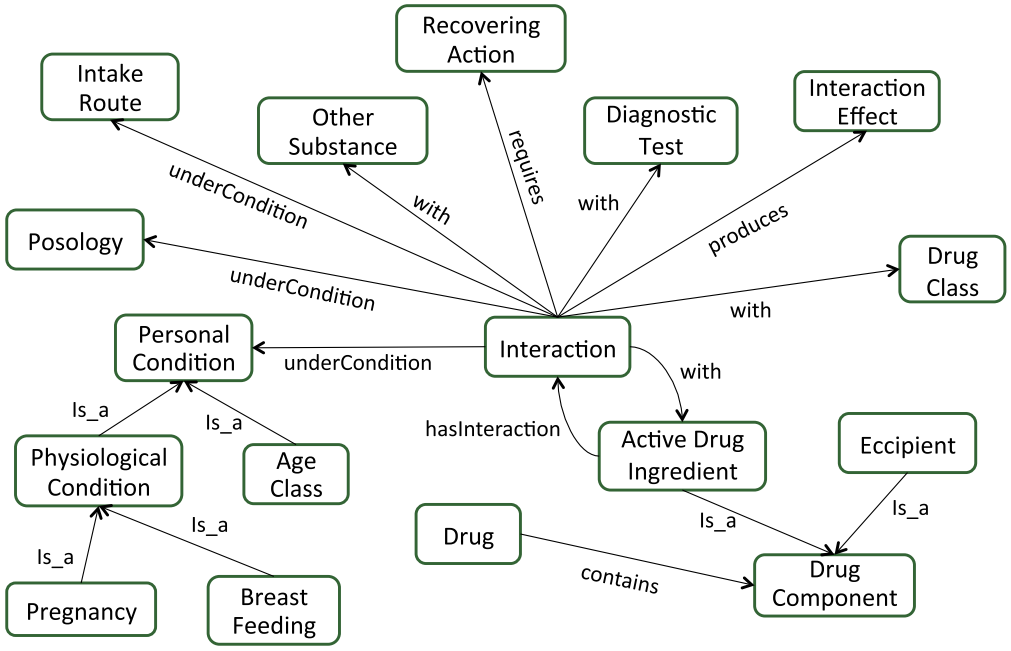

Fig. 1. Excerpt of the ontology concerned with drug-related interactions.

3.4. Features

The feature construction process aims at capturing the salient characteristics of each token in order to help the system to predict its semantic label. Feature definition is a critical stage regarding the success of feature based statistical models such as CRFs and SVMs.

Defining the feature function involves constructing a set of real-value features $b(x,i)$ of the observation x at position i in the sentence, to express some characteristics of empirical distribution of the training data that should also hold of the model distribution [33]. We will use binary features, an example of such a feature is:

$$b_j(x,i) = \begin{cases} 1: & \text{if the observation at position } i \text{ is} \\ & \text{the word } \texttt{pharmacokinetics} \\ 0: & \text{otherwise} \end{cases}$$

We implemented and employed a large variety of informative features. What follows is a report on the set of features we used in our experiments to discriminate between semantically interesting and uninteresting content.

3.4.1. Orthographical features

As a good starting point, our set of machine learning features consisted of the simplest and most obvious feature set: word identity features that is the vocabulary from the training data. Furthermore, we added features that indicate whether the current token is a digit, which is quite useful for identifying *Posology* entities.

3.4.2. Neighboring word features

Words preceding or following a target word may be useful for modeling the local context. It is clear that the more context words analyzed, the better and more precise the results become. However, widening the context window quickly leads to an explosion of the computational and statistical complexity. For our experiments, we estimated a suitable window size of $[-3,3]$. As an example, consider the following three features:

$$b_1(x,i) = \begin{cases} 1: & \text{if the observation at} \\ & \text{position } i \text{ is the word} \\ & \texttt{drugs} \\ 0: & \text{otherwise} \end{cases} \qquad b_2(x,i) = \begin{cases} 1: & \text{if the observation at} \\ & \text{position } i-1 \text{ is the} \\ & \text{word } \texttt{avoid} \\ 0: & \text{otherwise} \end{cases}$$

$$b_3(x,i) = \begin{cases} 1: & \text{if the observation at position } i+1 \text{ is} \\ & \text{the word } \texttt{association} \\ 0: & \text{otherwise} \end{cases}$$

Given the sequence "avoid drugs association", they would be active (i.e., equal to 1) for the middle token "drugs".

3.4.3. Prefix features

Some prefixes can provide good clues for classifying named entities. In particular, we identified a set of words that occur often with the same label; for example, Italian words starting with "effet-" (effect) or "farmacocinetic-" (pharmacokinetic) are usually *Interaction Effects*, those starting with "mg-" (mg) or "dos-" (dosage) or "giorn-" (day) have usually been tagged as *Posology*, and so on. Therefore, we also included some prefix features. These features help the system recognize informative substrings. However, short prefixes are too common to be of any help in classification. In our experience, the acceptable length for a prefix varies by words, and in many cases, the prefix coincides with the word root.

3.4.4. Punctuation features

Also notable are punctuation features, which contain some special punctuation in sentences. After browsing our corpus, we found that colon and brackets features might prove helpful. Given a medication in fact, colon is usually preceded by the interacting substance and followed by the explanation of the specific interaction effects. Additionally, round brackets denotes extra information regarding the words that follow. For each token, the punctuation features test whether it is preceded or followed by colon or parenthesis. These features have been used in conjunction with a token window equal to the sentence length. This means that punctuation features for token j contain predicates about the previous $j-n$ tokens and the following $j+m$ tokens, where n and m are the

distance between the current token and the beginning and the end of the sentence, respectively.

3.4.5. Dictionary features

In order to have this model benefit from domain-specific knowledge, we added semantic features. Farmadati DataBase is provided with a complete archive of active ingredients. We create a binary feature for each entry in the active ingredient archive. Every time a text token coincides with such an entry, the feature is active, indicating that the token describes an active drug ingredient. For dictionary entries that are multi-token (e.g., the active ingredient acetylsalicylic acid), all words are required to match in the input sequence.

3.4.6. Part of speech features

Finally, we supposed lexical information might be quite useful for identifying named entities. Thus, we included features that indicate the lexical function (also known as part of speech (POS)) of each token. We used TreeTagger, a POS tagger developed by Schmid [34], to provide POS information.

3.5. Discriminative structured prediction

In this section, we consider the problem of designing classification algorithms that learn a direct map from input vectors $x \in X$ to discrete output variables $y \in Y$, based on a training set of input–output pairs $(x_1, y_1), \ldots, (x_N, y_N) \in X \times Y$ drawn from some fixed, but unknown probability distribution. Unlike multi-class classification, where the output space consists of a finite set of scalar variables, in our case, the elements of Y are structured objects, in particular sequences of semantic labels, i.e., $\langle$*InteractionEffect*, *ActiveDrugIngredient*, $\ldots\rangle$. In order to deal with this kind of output, we avail ourselves of two discriminative classifiers for general structured and interdependent output variables: linear-chain CRFs and structured SVMs, a generalization of multi-class classifiers. Unlike generative models, discriminative classifiers learn a direct map from features x to the labels y. They are hence particularly successful in situations in which it is difficult to properly specify class-conditional densities. This is the case here in which we wish to incorporate a large variety of interdependent and long-range features of the data. The discriminative approach to classification therefore provides crucial modeling freedom.

3.5.1. Conditional random fields and support vector machines

Let X be a sequence of words in a text document, whose values are observed. Let Y be some sequence of semantic labels whose values the task requires the model to predict. CRFs and structured SVMs all learn linear discriminant functions F that acting on both X and Y, encode the interdependencies in the input–output space. Being x^* a novel observation sequence, we can derive a prediction of the label sequence y^* by maximizing F over the output variable. Hence the general form of our classifiers h is:

$$y^* = h_\theta(x^*) = \arg\max_{y \in Y} F(x^*, y, \theta) \tag{1}$$

where θ denotes a parameter vector to be estimated from training data. Based on this formulation, we can consider several learners that differ in how they choose model parameters.

The underlying idea of CRFs [35] is that of defining as discriminant function a direct model of the conditional probability $p(Y|X)$ distribution over label sequences given a particular observation sequence x, without assuming anything about the input distribution $p(X)$. Linear-chain CRFs define the conditional probability to be a normalized product of potential functions, each of the form:

$$\exp\left(\sum_j \lambda_j t_j(y_{i-1}, y_i, x, i) + \sum_k \mu_k s_k(y_i, x, i)\right) \tag{2}$$

where $t_j(y_{i-1}, y_i, x, i)$ is a transition feature function of the entire observation sequence and the labels at positions i and $i-1$ in the label sequence; $s_k(y_i, x, i)$ is a state feature function of the label at position i and the observation sequence; and $\theta = (\mu_k, \lambda_j)_{j,k}$. The parameters are set to maximize the conditional log-likelihood of labeled sequences in the training set. As a measure to control overfitting, we use a Gaussian regularization term.

Structured SVMs [36] minimize a particular trade-off between model complexity and empirical risk. Again, the discriminant function F takes the form $\langle\theta, \psi(x, y)\rangle$, where $\langle \cdot, \cdot \rangle$ denotes the inner product, ψ the feature vector function relating input and output, and θ are the model's parameters. To find θ, SVMs minimize the regularized empirical risk:

$$\min_\theta \|\theta\|^2 + C\sum_{i=1}^{N} \max_{\hat{y} \in Y} \{\Delta(y_i, \hat{y}) + \langle\theta, \psi(x_i, \hat{y})\rangle\} - \langle\theta, \psi(x_i, y_i)\rangle \tag{3}$$

The loss function $\Delta(y_i, \hat{y})$ indicate how far $\hat{y}$ is from the true output y_i.

The classifiers weigh the different features by estimating the associated parameters. Once the optimal parameter setting has been found, the models above can predict the label sequence for a previously unseen input. To achieve this, both classifiers solve the associated inference problem via Viterbi's algorithm [37]. Our system uses the MALLET [38] implementation of CRFs and an implementation by Bordes et al. [39] of SVMs.

4. Experimental evaluations

On the following lines, we present our evaluation method for the task of automatic recognition of drug-related entities.

4.1. Dataset

We created a corpus which consists of 100 manually annotated interaction sections of specialty medicines for human use. They have been extracted from monographs found in the Farmadati Italia Database, which were taken from the original SPCs drawn up by the producers. This database contains all the references about the medicines, the para-pharmaceutical, and the homeopathic products existing in Italy and the medical devices, which can be sold in pharmacies, for a total of about 800.000 recorded products. The interaction sections were derived using the BDF (Bancadati Federfarma) software, combined with a preprocessing algorithm. BDF software enabled the exporting of the monographs archive as an ASCII text format file. This file lists the monographs lines of each specialty medicines for human and veterinary use in the database. An alphanumeric code at the beginning of each line specifies the drug and the section it refers to. A preprocessing pass over the exported file allowed us to split it into different files, associated with the corresponding medicine and stating the different sections in each file. Moreover, our database may not be properly hyphenated due to some length constraints: we solved this problem using an Italian language lexicon [40].

4.2. Evaluation metrics

We measure the performance of our model on the individual labels using the standard evaluation metrics for machine learning algorithms: recall ((correct extractions)/(gold standard annotations)), precision ((correct extractions)/(total number of extractions)), and F_1-measure (the weighted harmonic mean of recall

and precision, with a weight set to 1 to give recall and precision the same importance) [41].

We report the results of our experiments showing recognition scores for the different labels and the overall performance evaluation of the two classifiers. Then, dealing with multi-label classification, we combined the performance results of the different labels following two principal approaches. Either we compute their arithmetic mean, giving equal weight to each of the labels (macro-averaged), or we compute the mean weighting each label by the number of times they occur in the dataset (micro-averaged). Macro-averaged metrics are often dominated by the performance on rare labels while micro-averaged metrics are dominated by the performance on frequent labels. The two ways of measuring performance are hence complementary, and both are informative.

4.3. Experimental setup

We randomly split the 100 interaction sections into two sets, one for training containing 60 sections and one for testing 40 sections. This amounts to a total of 796 input sentences for training and 457 input sentences for testing. We used cross-validation on the training sets to determine reasonable parameter settings of our models. For the linear SVMs, we found the regularization parameter $\lambda = 1$ to work well. All SVMs results in this paper have been produced using 10 passes through the entire training set. For the variance of the Gaussian regularizer of the CRFs [42], we used the value 10.

5. Results and discussion

The overall results of both the SVMs and the CRFs can be found in Table 1. In general, our experiments show that the classifiers, with carefully designed features, can identify drug interactions related information with a resulting overall accuracy of around 91%. The two models exhibit similar overall performance, without a clear superiority of one model over the other. Although the data might contain noise inherent to manual annotation, the learning algorithms reach good performance. Expressing the problem of content extraction in the described machine learning approach is therefore promising. However, though the rule-based approach is efficient, it still might be limited when processing complicated and highly variable text that conveys multiple kind of information.

Note that the values of the macro-averaged metrics are much lower than the micro-averaged one. The performance differences between the macro- and micro-averaged results suggest that some rare labels are often misclassified, as shown in their low recall in Table 2. Macro-averaged metrics, in fact, are often dominated by the performance on rare labels. Table 2 shows the performance results on the individual labels for all available features, in terms of precision recall and F_1-measure. Overall, labels whose training examples are scarce suffer from relatively low performance. It is the labels *DiagnosticTest* and *OtherSubstance* that are hardest to extract. On the other hand, some other labels such as *ClinicalCondition* and *IntakeRoute*, although rare, perform better. Such labels, in fact, can rely on a more precise definition, which is an important factor that contributes to the good performance. Also notable is the imbalance among the performance of the two models on the label *AgeClass*: on the one hand it is a rare label, on the other hand it is well defined in the texts. An explanation of this gap is left for future work.

Moreover, we investigate the influence of different feature groups (i.e., orthographical, neighboring word, prefix, POS, punctuation, and dictionary features) on the overall classification results. Table 3 looks at the performance when varying the employed set of feature, averaged over 20 trials. In each trial, the sentences in the new training set are sampled uniformly at random from the 1200 sentences in the original dataset. Each feature set differs only in the absence of a particular group, which is specified in the first column of Table 3. For all evaluations, we tested numerous settings for the variance of the Gaussian regularizer and found the value to work best. This is because each feature is, in general, associated with a parameter. The more parameters the model has, the higher its degree of freedom and the more likely it is to overfit. This means that in comparing of different feature groups, we should regularize differently. Particularly, the more features there are, the smaller

Table 1
Overall experimental results (in %) of CRFs and SVMs both including and not including the label *None*.

Model	Micro-average			Macro-average			Overall accuracy
	Precision	Recall	F_1-measure	Precision	Recall	F_1-measure	
CRF w/ *None*	91.27	91.35	91.13	91.98	74.91	80.83	91.34
CRF w/o *None*	89.97	81.97	85.56	91.98	72.77	79.51	81.98
SVM w/ *None*	91.46	91.53	91.41	91.33	80.32	84.99	91.52
SVM w/o *None*	83.63	86.25	86.86	91.19	78.78	84.07	83.63

Table 2
Performance results (in %) of the two classifiers on individual labels.

Label	N_{train}	N_{test}	CRF			SVM		
			Precision	Recall	F_1-measure	Precision	Recall	F_1-measure
ActiveDrug Ingredient	1379	711	97.98	95.64	96.80	97.85	95.92	96.88
Age Class	8	16	100	56.25	72.00	100	81.25	89.65
ClinicalCondition	33	69	100	66.67	80.00	100	72.46	84.03
DiagnosticTest	96	32	100	31.25	47.62	100	59.38	74.51
Drug Class	1223	939	87.75	76.25	81.60	88.10	78.81	83.19
IntakeRoute	39	22	90.48	86.36	88.37	90.00	81.82	85.71
InteractionEffect	1726	1136	87.58	81.95	84.67	86.82	82.92	84.83
None	12103	6899	91.96	96.38	94.12	92.65	95.77	94.18
Other Sub stance	97	80	76.47	65.00	70.27	71.23	65.00	67.97
PharmaceuticalForm	1	0	–	–	–	–	–	–
PhysiologicalCondition	3	0	–	–	–	–	–	–
Posology	418	213	93.17	89.67	91.39	96.48	90.14	93.20
Recovering Action	874	477	86.41	78.62	82.33	81.45	80.08	80.76

Table 3
Variation in performance (in %) for different features sets.

Feature set	Variance	CRF			SVM		
		Precision	Recall	F_1-measure	Precision	Recall	F_1-measure
All features	10	90.08	90.25	89.91	90.25	90.41	90.23
No POS features	10	89.86	90.00	89.70	90.03	90.13	89.97
No dictionary features	10	88.33	88.43	88.08	88.71	88.83	88.64
No number features	10	89.44	89.57	89.11	89.47	89.59	89.35
No prefix features	20	89.93	90.09	89.73	89.77	89.94	89.71
No punctuation features	20	89.63	89.77	89.40	89.58	89.74	89.51
No word identity features	100	88.071	88.36	87.97	88.15	88.39	88.13
No neighboring features	1000	85.88	86.22	85.80	85.83	86.23	85.45

the variance of the Gaussian prior needs to be. Figs. 2 and 3 depict the box-and-whisker plots of the results on different feature categories, showing minimal and maximal values as well as inter-quartile range (IQR) and median. We focus on the micro-averaged F_1-measure. These figures clearly illustrate that the absence of neighboring word features to a greater degree, and then word identity feature to a lesser, both worsen the micro-averaged F_1-measure with statistical significance. Not surprisingly, the neighboring word features have been shown to be the most beneficial ones when comparing the different feature sets, additionally because they represent the large majority of features. On the other hand, as expected, the effects of POS, number, punctuation,

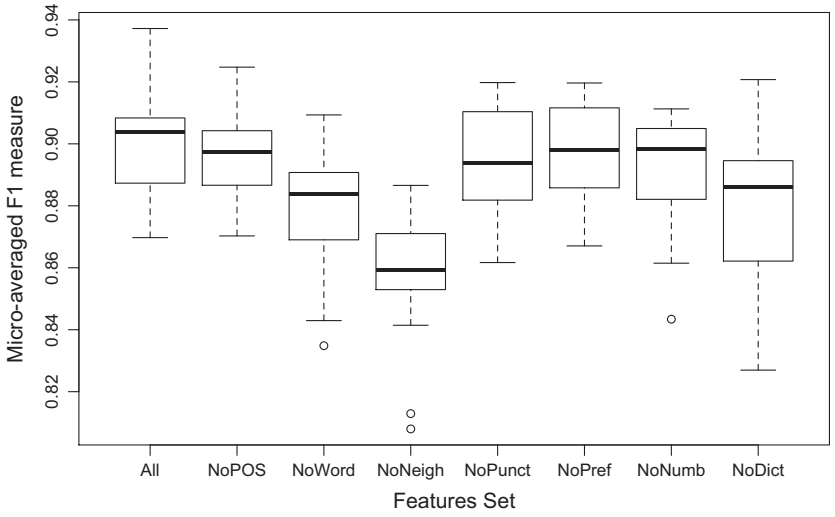

Fig. 2. Comparison of CRFs performance, in terms of micro-averaged F_1-measure, for different feature sets. On the horizontal axis, you can see the different feature sets which correspond exactly to those listed in the first column of Table 3. In this figure, they have been shortened due to length constrains.

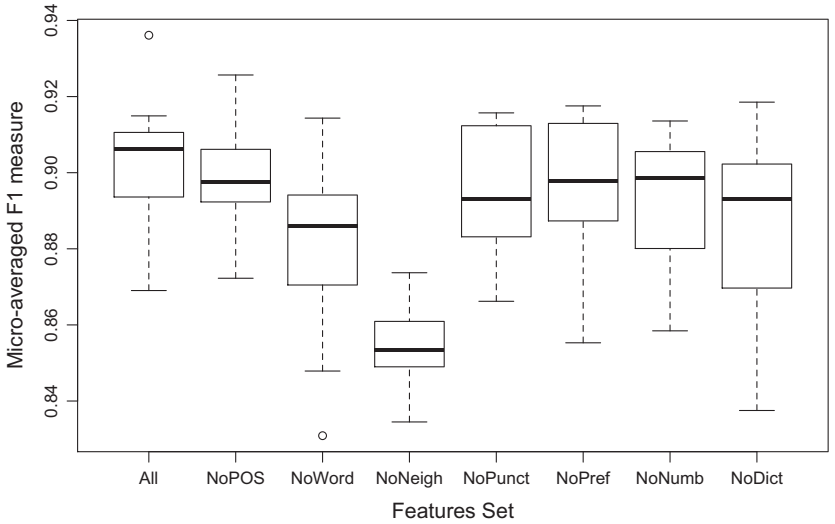

Fig. 3. Comparison of SVMs performance, in terms of micro-averaged F_1-measure, for different feature sets. On the horizontal axis, you can see the different feature sets which correspond exactly to those listed in the first column of Table 3. In this figure, they have been shortened due to length constrains.

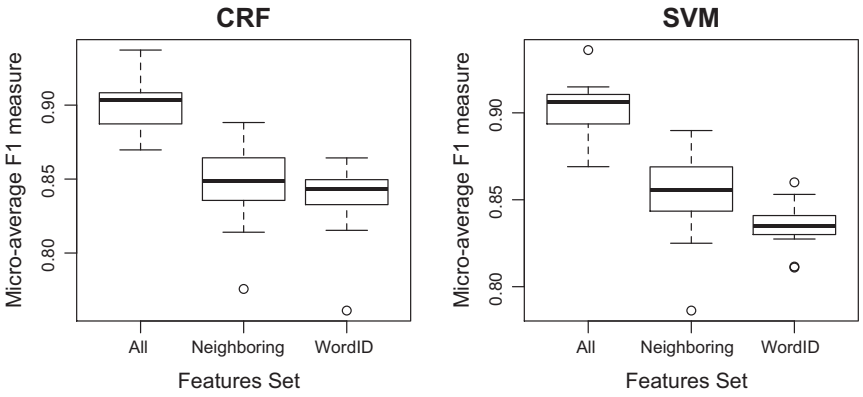

Fig. 4. CRFs and SVMs performance for varying feature sets.

prefixes, and dictionary features on these results are less strong, especially because there are only few of them. Finally, we studied the performances of our two models when using only word neighboring and word identity feature, respectively: Fig. 4 shows how the results considerably decrease with the above mentioned feature sets. Therefore, although POS, number, punctuation, prefixes, and dictionary features alone might not have a strong effect on the models performance, they contain information that is complementary, resulting in a synergy effects when merged.

The described results are comparable to existing systems. Gold et al. [24], for example, yielded a precision of 94.1% and a recall of 82.5%, Xu et al. [25] reported a *F*-measure over 90%. It should be noted, however, that such systems do not extract drug-interaction information, thus we cannot perform a fully accurate comparison. The performance of NER systems, in fact, usually changes when varying the type of information to be extracted. Within the i2b2 NLP challenge, for instance, among all types of medication information, duration and reasons were the most difficult to detect for all systems [26].

A more recent study from Segura-Bedmar et al. [31] has focused on the same task of drug-interaction extraction. They follow a similar SVMs-based approach and yet rely on a corpus of documents from the DrugBank database. A second aspect is that this approach differs from our own since it detects only whether or not an interaction occurs between a given pair of drugs in a sentence and does not consider more specific information about each interaction. Although this work is hardly comparable with our own, we nevertheless report results: their best method achieves a F_1-measure of 60.01%. Our experimental results compare favorably, since they have a F_1-measure of about 91%.

6. Conclusions

We have shown that is possible to perform the task of information extraction from SPCs using supervised machine learning techniques. Although we have focused on drug interactions, the encouraging results and the adaptability of the approach we adopted means that our system has general significance for the extraction of detailed information about drugs (drug targets, contraindications, side effects, etc.).

Acknowledgments

We gratefully thank Patrick Gallinari and Alex Spengler from the Laboratoire d'Informatique de Paris 6 of the Université Pierre et Marie Curie, for their helpful feedback on the system design and implementation.

References

[1] Bates DW, Cullen DJ, Laird N, Petersen LA, Small SD, Servi D, et al. Incidence of adverse drug events and potential adverse drug events implications for prevention. ADE prevention study group. J Am Med Assoc 1995;274:29–34.
[2] Institute of Medicine, editor. Preventing medication errors. Washington: The National Academics Press; 2007.
[3] Bates DW, Boyle DL, Vander Vliet M, Schneider J, Leape LL. Relationship between medication errors and adverse drug events. J Gen Intern Med 1995;10:100–205.
[4] Hook J, Cusack C. Ambulatory computerized provider order entry (CPOE): findings from the AHRQ Health IT portfolio. Center for information technology leadership. AHRQ National Resource Center for Health Information Technology, Rockville: Agency for Healthcare Research and Quality, AHRQ publication no. 08-0063-ef edition; 2008.
[5] Kaushal R, Shojania KG, Bates DW. Effects of computerized physician order entry and clinical decision support systems on medication safety: a systematic review. Arch Intern Med 2003;163:1409–16.
[6] Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. J Am Med Assoc 2005;293:1223–38.
[7] Eslami S, Abu-Hanna A, De Keizer NF. Evaluation of outpatient computerized physician medication order entry systems: a systematic review. J Am Med Inform Assoc 2007;14:400–6.
[8] Shamliyan TA, Duval S, Du J, Kane RL. Just what the doctor ordered. review of the evidence of the impact of computerized physician order entry system on medication errors. Health Serv Res 2008;43:32–53.
[9] Wolfstadt JI, Gurwitz JH, Field TS, Lee M, Kalkar S, Wu W, et al. The effect of computerized physician order entry with clinical decision support on the rates of adverse drug events: a systematic review. J Gen Intern Med 2008;23:451–8.
[10] Ammenwerth E, Schnell-Inderst P, Machan C, Siebert U. The effect of electronic prescribing on medication errors and adverse drug events: a systematic review. J Am Med Inform Assoc 2008;15:585–600.
[11] Evans RS, Pestotnik SL, Classen DC, Clemmer TP, Weaver LK, Orme JJ, et al. A computer-assisted management program for antibiotics and other antiinfective agents. New Engl J Med 1998;338:232–8.
[12] Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996;35:285–301.
[13] Friedman C, Hripcsak G. Natural language processing and its future in medicine. Acad Med 1999;74:890–5.
[14] Sager N, Friedman C, Chi E. The analysis and processing of clinical narrative. In: Salamon R, Blum B, Jorgensen M, editors. Proc fifth conference on medical informatics. Elsevier Science Publishers; 1986. p. 1101–5.
[15] Friedman C, Johnson SB, Forman B, Stanner J. Architectural requirements for a multipurpose natural language processor in the clinical environment. In: Proc annu symp comput appl med care. p. 347–51.
[16] Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med 1998:1–7.
[17] Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports Radiology 1990;174:543–48.
[18] McCarry AT, Sponsler JL, Brylawski B, Browne AC. The role of a lexical knowledge in biomedical text understanding. In: Proc annu symp comput appl med care. p. 103–4.
[19] Ananiadou S, mcNaught J. Text mining for biology and biomedicine. Artech House, Inc.; 2006.
[20] Levin MA, Krol M, Doshi AM, Reich DL. Extraction and mapping of drug names from free text to a standardized nomenclature. In: AMIA annu symp proc. p. 438–42.
[21] Sirohi E, Peissig P. Study of effect of drug lexicons on medication extraction from electronic medical records. In: Pac symp biocomput, vol. 10. p. 308–18.

[22] Evans DA, Brownlowt ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. In: Proc AMIA annu fall symp. p. 388–92.
[23] Shah AD, Martinez C. An algorithm to derive a numerical daily dose from unstructured text dosage instructions. Pharmacoepidemiol Drug Saf 2006;15:161–6.
[24] Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. In: AMIA annu symp proc. p. 237–41.
[25] Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. Medex: a medication information extraction system for clinical narratives. J Am Med Inform Assoc 2010;17:19–24.
[26] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17:514–8.
[27] Doan S, Bastarache L, Klimkowski S, Denny J, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. J Am Med Inform Assoc 2010;17:528–31.
[28] Tikk D, Solt I. Improving textual medication extraction using combined conditional random fields and rule-based systems. J Am Med Inform Assoc 2010;17:540–4.
[29] Meystre S, Thibault J, Shen S, Hurdle J, South B. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. J Am Med Inform Assoc 2010;17:559–62.
[30] Pereira S, Plaisantin B, Korchia B, Rozanes N, Serrot E, Joubert M, et al. Automatic construction of dictionaries, application to product characteristics indexing. In: Workshop on advances in bio text mining.
[31] Segura-Bedmar I, Martínez P, de Pablo-Sánchez C. Using a shallow linguistic kernel for drug–drug interactions extraction. J Biomed Inform 2011. in press.
[32] Rubrichi S, Leonardi G, Quaglini S. A drug ontology as a basis for safe therapeutic decision. In: Proc second italian national conference of bioengineering, Patron; 2010.
[33] Wallach HM. Conditional random fields: an introduction. Technical report, MS-CIS-04-21. University of Pennsylvania; 2004.
[34] Schmid H. Treetagger a language independent part of speech tagger. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger>, ????.
[35] Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proc international conference on machine learning, vol. 18. p. 282–289.
[36] Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. J Mach Learn Res 2005;6:1453–84.
[37] Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory 1967;13:260–9.
[38] McCallum A. Mallet: a machine learning for language toolkit. Technical report; 2002. <http://mallet.cs.umass.edu>.
[39] Bordes A, Usunier N, Bottou L. Sequence labelling SVMs trained in one pass. In: Proc ECML PKDD. Springer; 2008. p. 146–61.
[40] Zanchetta E, Baroni M. Morph-it!: a free corpus-based morphological resource for the italian language. In: Proc corpus linguistics conference.
[41] Van Rijsbergen CJ. Information Retrieval. Department of Computer Science. University of Glasgow; 1979.
[42] Sutton C. GRMM: graphical models in mallet. <http://mallet.cs.umass.edu/grmm/>; 2006.